HYPOTHESIS & PREDICTION

Hypothesis: Host order is a significant driver of whether a mammal species is more likely to be affected by a segmented virus. Segmented viruses have genomes that are divided into multiple segments, allowing for detailed control over gene expression and protein production. This segmentation enables the virus to respond dynamically to host environments.

The relationship between segmentation and host order likely stems from the fact that host species with complex genomes or those requiring more specialized interactions with the virus may drive the evolution of segmented viral genomes to accommodate these needs. However, other predictors should also be considered. These should be selected based on their biological relevance to viral infectability. Host order, combined with these predictors, provides a comprehensive analysis for understanding susceptibility to segmented viruses.

Some predictors we are planning to use are:

1. Genome length
2. Enveloped or non-enveloped
3. DNA or RNA

The final set of predictor variables will be decided upon further literature review.

Prediction: If host order is a crucial determinant of a mammal species' susceptibility to segmented viruses, then a model that includes host order alongside other predictors will be more compatible with data compared to a model with only the other predictors. Specifically, the model with host order will have the lowest AIC, indicating better fit and predictive accuracy compared to the model without host order.

Such superiority of the model with host order would indicate that host order accounts for a big portion of the variation in susceptibility compared to other predictors. This further suggests that host order captures essential biological factors that influence how segmented viruses infect different mammal species. For example, host order may represent shared evolutionary traits that makes certain orders more or less susceptible to segmented viruses or segmented viruses may have evolved throughout time to better suit the biology of certain host orders.

DATA DESCRIPTION

The data we will use for this analysis was retrieved from the Clover database, a resource developed by a team of researchers led by Rory Gibb and collaborators. Clover is a comprehensive database on host-virus association from four major datasets: EcoHealth Alliance's HP3, the University of Georgia's GMPD2, the University of Liverpool's EID2 and a dataset gathered by Shaw et al. These datasets draw on information from experimental studies, field observations, and literature reviews to map virus-host interactions. Clover has also been edited to keep the data's taxonomy consistent with NCBI. We will focus on segmented viruses, particularly those that are transmitted cross-species (zoonotic viruses). Specifically, Mammal_viruses-Associations and the HP3_virus dataset were selected from the Clover database.

The Mammal_Viruses_Associations dataset provides information on mammalian host orders and associated viruses. It tells us how to group each host species based on host order, along with the viruses that affect each species. The HP3_virus dataset provides virus-level information, such as whether a virus has a segmented genome, virus family, average genome size, etc.

Initially, we sampled 1000 rows for testing purposes during code development, in the final analysis, we will be sampling the whole dataset. To prepare our data, virus names from both datasets will be standardized by replacing spaces with underscores and converting them to lowercase to ensure consistency. The data frame will be subsetted based on columns of interest, which will eventually be predictors for our model. Then, our datasets will be merged based on virus names, with the resulting dataset further manipulated by removing rows with missing values. We will also generate the number of segmented viruses and total viruses for each host order to understand the proportion of segmented viruses across host orders.

DATA ANALYSIS

We predict that host order affects whether a mammal species has a greater proportion of infection by segmented viruses relative to non-segmented viruses. Two generalized linear models (GLMs) following a binomial distribution will be tested against the data. Both models will regress the proportion of viruses that are segmented against multiple predictor variables—the predictors will be specified later. Only one of the two models will have host order as a predictor variable. Then, both models will be evaluated on their compatibility with the data. For our prediction to be correct, the GLM including the host order predictor will fit the data better and have lower AIC.

Our data follows a binomial distribution because we can treat segmentation as 'success'—i.e. a segmented virus is successful and equals 1, while a non-segmented virus is unsuccessful and equals 0. Since we are interested in host order, rather than host species, we can use each mammal species-virus association as a trial.